УДК 004.942:612.2

МНОГОКРИТЕРИАЛЬНЫЙ АЛГОРИТМ ШАГОВОЙ РЕГРЕССИИ

Настенко Е.А., д.б.н., к.т.н.
nastenko.e@gmail.com
Кафедра биомедицинской кибернетики
Национального технического университета
«Киевский политехнический институт имени Игоря Сикорского»,
г. Киев, Украина
Павлов В.А,, доц., к.т.н.
pavlov.vladimir264@gmail.com
Кафедра биомедицинской кибернетики
Национального технического университета
«Киевский политехнический институт имени Игоря Сикорского»,
г. Киев, Украина
Бойко А.Л., доц., к.п.н.
annaanna2612@gmail.com
Кафедра физического воспитания
Национального технического университета
«Киевский политехнический институт имени Игоря Сикорского»,
г. Киев, Украина
Носовец Е.К., к.т.н.
o.nosovets@gmail.com
Кафедра биомедицинской кибернетики
Национального технического университета
«Киевский политехнический институт имени Игоря Сикорского»,
г. Киев, Украина

Реферат *– В работе решается задача структурно-параметрического синтеза модели множественной линейной регрессии в условиях частичной мультиколлинеарности входных переменных. Алгоритм осуществляет исключение переменных по оценке мультиколлинеарности методом Фаррара-Глобера. Оптимизация параметров шагового алгоритма осуществляется в соответствии с внешним критерием - нормированная относительная среднеквадратичная ошибка на проверочной выборке данных. Рассмотрен пример моделирования эффективности функционального состояния дыхательной системы пациента по величине потребления кислорода. Сравнение результатов стандартного Stepwise и предложенного алгоритмов показало преимущество последнего на экзаменационной выборке данных.*

Ключевые слова *– принципы самоорганизации, шаговый алгоритм, многомерная линейная регрессия, мультиколлинеарность, метод Фаррара-Глобера, внешний критерий, функциональное состояние, дыхательная система.*

I. ВВЕДЕНИЕ

Для получения моделей множественной регрессии наиболее часто применяют шаговые структурно-параметрические методы: шаговая регрессия Stepwise, метод наименьших углов, ступенчатая регрессия, последовательное добавление признаков с их ортогонализацией, Лассо [1]. Каждый из них реализует последовательное добавление или удаление признаков по определенному критерию. Общим недостатком существующих процедур является отсутствие гарантии, что истинная (по модельным экспериментам) структура модели будет найдена. Кроме того, переобученные (с высокой точностью на обучающей выборке) модели при наличии частично мультиколлинеарных аргументов невозможно применять на новых данных. Однако, именно хорошее качество найденной модели на "свежих" точках является действительной задачей структурно-параметрического синтеза.

Явление частичной мультиколлинеарности входных аргументов приводит к росту дисперсии оценки параметров регрессионной модели и затрудняет объяснение влияния входных переменных на зависимую переменную. Как следствие, существенная мультиколлинеарность приводит к

невозможности оценки такой моделью значений зависимой переменной. Стандартные шаговые процедуры отбора факторов не оценивают степень мультиколлинеарности отобранных аргументов и не решают указанную проблему. Поэтому, разработка структурно-параметрических методов множественной регрессии с учетом ограничения частичной мультиколлинеарности является актуальной.

В настоящей работе рассматривается многокритериальный подход, обеспечивающий возможность оптимизации параметров алгоритма на принципах самоорганизации для получения модели, удовлетворяющей требованиям наилучшей точности и ограничениям частичной мультиколлинеарности.

II. ЦЕЛЬ ИССЛЕДОВАНИЙ

Целью работы является разработка алгоритма построения регрессионной модели, оптимизирующей значение нормированной относительной среднеквадратичной ошибки на тестовой выборке данных с учетом ограничения частичной мультиколлинеарности входных переменных.

III. ПОСТАНОВКА ЗАДАЧИ

Задана матрица входных наблюдений $x \in R^M$ и вектор зависимой переменной y:

$$\begin{pmatrix} x_{11} & \dots & x_{1M} & y_1 \\ x_{21} & \dots & x_{2M} & y_2 \\ \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nM} & y_n \end{pmatrix},$$

где n - число наблюдений, M - число факторов, из которых необходимо выбрать m наилучших объясняющих аргументов в модель.

Необходимо предложить алгоритм структурно – параметрического синтеза модели оптимальной структуры:

$$Y = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + .. + + \beta_m x_{i_m}, \quad \text{где} \quad m < M.$$

Оптимальность структуры модели понимается в смысле достижения наилучшего значения нормированной относительной среднеквадратичной ошибки на тестовой выборке данных с ограничением по параметру мультиколлинеарности набора входящих в модель аргументов.

IV. РЕЗУЛЬТАТЫ

Для решения указанной выше проблемы за основу взят алгоритм шаговой регрессии Stepwise. В классическом варианте алгоритма необходимость включения и исключения переменных оценивается критерием Фишера как отношение изменения дисперсии модели к ее среднеквадратичной ошибке [2]. Как следствие повышается точность модели на обучающей выборке и игнорируется возможная мультиколлинеарность включенных переменных. В таком случае оценки параметров модели для разных выборок могут сильно отличаться несмотря на то, что выборки однородны. Ниже предлагается решение данной проблемы на основе принципов самоорганизации: оптимизируются параметры алгоритма, обеспечивающие требования по точности и частичной мультиколлинеарности в соответствии со значением внешнего критерия - нормированной относительной среднеквадратичной ошибке на тестовой выборке.

Модификация метода классической шаговой регрессии заключается в следующем: включение аргументов осуществляется по критерию Фишера, как отношение изменения дисперсии модели к ее среднеквадратичной ошибке [2]; исключение осуществляется, используя модификацию метода Фаррара-Глобера [3]**:** по величине критерия Фишера, характеризующего дисперсию оценки параметров модели, если расчетное значение статистики, характеризующей величину детерминанта корреляционной матрицы, будет менее ее заданного значения. Оптимизация пороговых значений параметров происходит по значению внешнего критерия [4]**,** как величины нормированной относительной среднеквадратичной ошибки на тестовой выборке данных.

Далее рассмотрим основные этапы алгоритма:

1. Расширяем матрицу входных переменных X.

2. Выборку наблюдений делим на обучающую, проверочную и экзаменационную в заданном соотношении при сохранении однородности по дисперсии зависимой переменной.

Далее синтез модели проводится на рабочей выборке, а конечная оценка качества и сравнение результатов с базовым пошаговым алгоритмом Stepwise осуществляется на экзаменационной выборке.

3. Задается первоначальное (следующее текущее) значение порогов $F_{вкл}$ критерия Фишера для включения аргумента и $f_{искл}$ преобразованной статистики хи-квадрат, предложенной в методе Фаррара-Глобера для исключения аргумента, из заданных пределов перебора их значений.

4. Модифицированный шаговый алгоритм для фиксированных значений $F_{вкл}$ и $f_{искл}$ состоит из нескольких этапов:

4.1. Включение переменной в модель.

Процедура использует F критерий Фишера (1) для отношения изменения дисперсии модели к ее среднеквадратичной ошибке.

$$F_i = \frac{SSR_{prev+x_i} - SSR_{prev}}{MSR_{prev+x_i}}, \qquad (1)$$

$$MSR_{prev+x_i} = \frac{SSR_{prev+x_i}}{d}, \qquad (2)$$

$$d = n - m - 2, \qquad (3)$$

$$SSR = \sum_{i=1}^{n} (\overline{Y}_i - Y_i)^2, \qquad (4)$$

где MSR (2) – остаточное среднее квадрата ошибки модели; $SSR(4)$ – остаточная сумма квадратов (ошибок); $d(3)$ – число степеней свободы остатков (или ошибок) $i = 1...k_i$, k_i – количество факторов-претендентов, ранее не включенных в модель; x_i – вводимый предиктор; Y_i – табличные (исходные) значение переменной; $\overline{Y}_i$ – значение регрессионной модели; n – количество наблюдений в выборке; m – количество переменных в модели, $prev$ – модель, полученная на предыдущей итерации.

4.2. Рассчитанные значения F-критериев сравниваются с пороговыми $F_{вкл}$. Если $F_k > F_{вкл}$, то принимаем гипотезу о целесообразности включения переменной x_i в модель [2].

5. Исключение переменной из модели.

5.1. Механизм исключения переменной использует модификацию подхода Фаррара-Глобера [3]. Для этого рассчитываем оценку присутствия частичной мультиколлинеарности в векторе входных переменных модели:

$$f = -\left[n - 1 - \frac{2m+5}{6} \right] \log(1 - \Delta R), \qquad (5)$$

где ΔR – определитель корреляционной матрицы:

$$R = \begin{pmatrix} r_{x_1x_1} & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{x_2x_1} & r_{x_2x_2} & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_mx_1} & r_{x_mx_2} & \dots & r_{x_mx_m} \end{pmatrix}$$

5.2. Рассчитанное значение f сравнивается с пороговым $f_{искл}$. Если $f < f_{искл}$, то принимается гипотеза о присутствии в наборе признаков мультиколлинеарности и рассчитываются F_k критерии:

$$F_k = \frac{(d_{kk} - 1)(n - m)}{m - 1}, \qquad (5)$$

где d_{kk} – диагональные элементы матрицы, обратной матрице корреляций:

$$R^{-1} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1m} \\ d_{21} & d_{22} & \dots & d_{2m} \\ \dots & \dots & \dots & \dots \\ d_{m1} & d_{m2} & \dots & d_{mm} \end{pmatrix}.$$

5.3. Исключение k-й переменной применяется из условия максимума критерия $F_{k\max}$.

Для формируемой модели расчет вектора параметров осуществляется в соответствии с выражением $\overline{\beta} = (X^T X)^{-1} X^T Y$ для выбранной структуры X на обучающей выборке [5].

5.4. Если предикторы для включения не исчерпаны, переходим к п.5.1, если исчерпаны - к п. 6.

6. Рассчитывается значение внешнего критерия - коэффициент детерминации (4) на проверочной выборке данных:

$$r^2 = 1 - \frac{\sum_{i=1}^{n_{тест}} (y_i - \bar{y}_i)^2}{\sum_{i=1}^{n_{тест}} (y_i - \bar{y})^2}, \qquad (7)$$

где $y_i, \bar{y}_i$ – значение табличных и модельных точек проверочной выборки данных, $\bar{y}$ – среднее значение табличных точек проверочной выборки данных [5].

7. Если значения сетки $F_{вкл}$ и $f_{искл}$ не исчерпаны, переходим к п.3, иначе – к п.8.

8. Выбирается модель с набором $x_{i_1},..,x_{i_m}$ и параметры $F*_{вкл}, f*_{искл}$, на которых достигнут минимум внешнего критерия.

9. Возвращаемся к п.4. и строим модель с найденными в п.8 оптимальными параметрами $F*_{вкл}$ и $f*_{искл}$ на рабочей выборке.

10. Окончательный выбор модели из полученных по п.8 и п.9 осуществляется по значению нормированной относительной среднеквадратичной ошибке на экзаменационной выборке.

V. СРАВНЕНИЕ РЕЗУЛЬТАТОВ БАЗОВОГО И ПРЕДЛОЖЕННОГО АЛГОРИТМОВ

Для сравнения моделей, построенных по Stepwise с предложенной выше версией алгоритма шаговой регрессии с оптимизацией параметров были взяты анонимизированные данные, полученные в НИССХ им. Н.Н. Амосова. Выборка содержит 140 наблюдений: зависимую переменную и 5 входных переменных.

Цель задачи - получить модель эффективности функционального состояния дыхательной системы, предсказываемую по величине потребления кислорода организмом пациента (зависимая переменная).

Входные аргументы:

- x_1 – содержание кислорода в артериальной крови (в объемных процентах);
- x_2 – содержание кислорода в венозной крови (в объемных процентах);
- x_3 – центральное венозное давление (в мм. рт. ст.);
- x_4 – температура пациента (в °C);
- x_5 – нормированный ток крови (производительность аппарата искусственного кровообращения).

Кроме имеющихся факторов рассмотрим обобщенные переменные: x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_2x_3, x_2x_4, x_2x_5, x_3x_4, x_3x_5, x_4x_5. Выборка наблюдений была разделена на обучающую, проверочную и экзаменационную (по однородности дисперсии зависимой переменой) в количествах $80-40-10$ точек. Для увеличения сложности модели и демонстрации преимущества предложенного алгоритма, выполнено функциональное преобразование выхода как $Y=y^3$.

Построение модели для классического Stepwise и предложенного алгоритмов производилось на рабочей выборке (рабочая = обучающая + проверочная).

Уравнение модели, полученной по Stepwise, имеет вид:

$$y = -85811.3 - 462790 \cdot x_1 + 571607 \cdot x_2 - 9877.65 \cdot x_5 + 399408 \cdot x_1x_3 + 2704.94 \cdot x_1x_5 - 471460 \cdot x_2x_3 - 2727.21 \cdot x_2x_5$$

Значение нормированной относительной среднеквадратичной ошибки на экзамене: $\Delta_1 = 0.61$

Уравнение регрессии, полученное предложенным многокритериальным алгоритмом, имеет вид:

$$y = -382976 - 441131 \cdot x_1 + 600543 \cdot x_2 - 18523.9 \cdot x_4 + 471003 \cdot x_1x_3 - 572233 \cdot x_2x_3 + 5339.91 \cdot x_3x_4$$

Модель получена для оптимальных значений $F*_{вкл} = 3{,}84$, $f*_{искл} = 0{,}002$.

Значение нормированной относительной среднеквадратичной ошибки на экзамене: $\Delta_2 = 0{,}543$.

ВЫВОДЫ

В работе предложен многокритериальный алгоритм синтеза линейной множественной регрессии с учетом ограничения мультиколлинеарности входных переменных. Для оптимизации значений параметров алгоритма использован внешний критерий – нормированная относительная среднеквадратичная ошибка на тестовой выборке. Рассмотрен пример моделирования эффективности функционального состояния

дыхательной системы пациента по величине потребления кислорода. Сравнение результатов стандартного шагового алгоритма Stepwise с предложенным многокритериальным алгоритмом показало улучшение качества модели на 6.5% на экзаменационной выборке данных. Улучшение показателя на экзаменационной выборке указывает на повышение устойчивости данной модели за счет выбора решения с ограниченным уровнем мультиколлинеарности и оптимальным значением порогов отбора аргументов в модель.

ПЕРЕЧЕНЬ ССЫЛОК

[1] Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. - с. 15-37

[2] Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ, Мир, 1982. – с. 141-222.

[3] Фаррар Д.Е., Глаубер Р.Р. Мультиколлинеарность в регрессионном анализе: обзор эконометрики и статистики, часть 2, 49 том, 1 номер, Февраль, 1967, стр. 92-107

[4] Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. Киев: «Наукова думка», 1985, - 216 с

[5] Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. — 2-е изд. — М., 1962.

УДК: 004.942:612.2

БАГАТОКРИТЕРІАЛЬНИЙ АЛГОРИТМ КРОКОВОЇ РЕГРЕСІЇ

Настенко Є.А., д.б.н., к.т.н.
nastenko.e@gmail.com
Кафедра біомедичної кібернетики
Національного технічного університету
«Київський політехнічний інститут імені Ігоря Сікорського»,
м.Київ, Україна
Павлов В.А,, доц., к.т.н.
pavlov.vladimir264@gmail.com
Кафедра біомедичної кібернетики
Національного технічного університету
«Київський політехнічний інститут імені Ігоря Сікорського»,
м.Київ, Україна
Бойко Г.Л., доц., к.п.н.
annaanna2612@gmail.com
Кафедра фізичного виховання
Національного технічного університету
«Київський політехнічний інститут імені Ігоря Сікорського»,
м.Київ, Україна
Носовець О.К., к.т.н.
o.nosovets@gmail.com
Кафедра біомедичної кібернетики
Національного технічного університету
«Київський політехнічний інститут імені Ігоря Сікорського»,
м.Київ, Україна

Анотація. *У роботі вирішується завдання структурно-параметричного синтезу моделі множинної лінійної регресії в умовах часткової мультиколінеарності вхідних змінних. Алгоритм здійснює виключення змінних за оцінкою мультиколінеарності методом Фаррара-Глобера. Оптимізація параметрів крокового алгоритму здійснюється відповідно до зовнішнього критерію - нормированій відносній середньоквадратичній помилці на перевірочній вибірці даних. Розглянуто приклад моделювання ефективності стану дихальної системи пацієнта за величиною спожитого кисню. Порівняння результатів стандартного Stepwise та запропонованого алгоритмів показало перевагу останнього на екзаменаційній вибірці даних.*

Ключові слова*: принципи самоорганізації, крековий алгоритм, багатовимірна лінійна регресія, мультиколінеарність, метод Фаррара-Глобера, зовнішній критерій, функціональний стан, дихальна система.*

UDC 004.942:612.2

MULTI-CRITERION STEP-REGRESSION ALGORITHM

Nastenko E.A., Doctor of Science
nastenko.e@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
Kiev, Ukraine
Pavlov V.A., Associate Professor, Ph.D.
pavlov.vladimir264@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
Kiev, Ukraine
Boyko A.L., Associate Professor, Ph.D.
annaanna2612@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
Kiev, Ukraine
Nosovets O.K., Ph.D.
o.nosovets@gmail.com
Department of Biomedical Cybernetics
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”,
Kiev, Ukraine

Abstract – *The problem of structural-parametric synthesis of multiple linear regression models on the input variables in conditions of partial multicollinearity is solved in the work. The partial multicollinearity phenomenon of input arguments conduces to a variance estimates increase of the regression model parameters, and makes difficult to explain the impact of input variables to the dependent variable. Substantial multicollinearity conduces to impossibility of output estimation model. Classical stepwise procedures of the factors selection do not solve the problem of multicollinearity. Thus, the design of structural-parametric multiple regression methods considering limitation of partial multicollinearity are relevant.*

Keywords – *Principles of self-stepping algorithm, multivariate linear regression, multicollinearity method by Farrar-Glauber, external criterion, functional status, respiratory system..*